Problem Set 9
"Two-Variable Regression: Gauss-Markov Properties"

(9.1) [Error term with nonzero mean] Consider the linear model

(i) $y = \beta_1 + \beta_2 x + \varepsilon,$

where β_1 and β_2 are unknown parameters and ε is a random error term, but suppose $E(\varepsilon|x) = 5$ instead of zero. (The other Gauss-Markov properties still hold.) Econometrician #1 claims "Model (i) cannot be estimated by least-squares because the mean of the random error term is not zero." Econometrician #2 claims "Yes it can. Just estimate the following model by least squares:

(ii) $y = \gamma_1 + \gamma_2 x + \delta,$

where δ is a new error term assumed to have mean zero, and then make slight adjustments to the estimated coefficients" Back up Econometrician #2's claim as follows.

a. Let $\delta = -5 + \varepsilon$ and prove that $E(\delta) = 0$, using the rules for the expectation operator (E).
b. Substitute $\delta = -5 + \varepsilon$ into equation (ii) and rearrange to group the constants together.
c. What adjustment, if any, must be made to the estimate of γ_1 to get an estimate of β_1?
d. What adjustment, if any, must be made to the estimate of γ_2 to get an estimate of β_2?

(9.2) [Fundamental assumptions] Suppose we have a dataset of households in various cities of the U.S. We want to estimate the household demand relationship for gasoline using the quantity demanded by each household in a recent month and the average price per gallon paid by that household during that month. So we estimate $Q = \beta_1 + \beta_2 P + \varepsilon$, where Q denotes the quantity demanded and P denotes the price of the good. Our goal is to estimate the effect of price on quantity demanded, *ceteris paribus*.

a. Household income I likely influences the quantity demanded of gasoline, though it is omitted from this equation. Would you expect income to have a positive or negative effect on the quantity demanded?
b. Suppose income is positively correlated with price—that is, households with higher incomes on average must pay more for gasoline. If income is omitted from the regression equation—perhaps for lack of data—will the least-squares estimate of β_2 be biased up (toward zero), biased down (too negative), or unbiased? Justify your answer.

(9.3) [Fundamental assumptions] Suppose you want to estimate an equation relating average graduation rates from high school (grad) to average family income (inc) using a sample of metropolitan areas: $grad_i = \beta_1 + \beta_2\, inc_i + \varepsilon_i$. Consider the effects of parents' education. You do not have data on the average education level of parents, so it must be relegated to the error term. But you need to think about how parents' education might affect your estimate of β_2.

a. Do you think parents' education is likely to be positively correlated with average family income, negatively correlated, or uncorrelated? Why?
b. Do you think parents' education is likely to have a direct positive effect on high school graduation rates, a negative effect, or no effect? Why?
c. Assume that you want to estimate the *ceteris paribus* effect of average family income on graduation rates, because you want to know whether government programs that raise income will also raise graduation rates. Will the least squares estimator of β_2 be biased up, biased down, or unbiased? Justify your answer.

(9.4) [Fundamental assumptions] Consider the two-variable model *without* an intercept, that is $y = \beta x + \varepsilon$. Notice that there is only one β parameter for this model—the slope. There is no intercept because the true population regression function is here assumed to pass through the origin. It can be shown that the least-squares estimator of β is given by $\hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$. where the sums run from i=1 to n. Assume that $E(\varepsilon_i \mid x_i) = 0$. This means that you can treat the x_i as fixed numbers, not random variables, and that $E(\varepsilon_i) = 0$.

a. Prove that $\hat{\beta} = \beta + \dfrac{\sum x_i \varepsilon_i}{\sum x_i^2}$. Justify each step of your proof.
b. Prove that the least-squares estimator of β is unbiased, that is $E(\hat{\beta}) = \beta$. Justify each step of your proof.
c. Further assume that $Var(\varepsilon_i \mid x_i) = \sigma^2$ and that $Cov(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$. Derive the formula for the variance of the least-squares estimator $Var(\hat{\beta}) = E\left(\left(\hat{\beta} - \beta\right)^2\right)$. Justify each step of your proof. [Hint: Your formula will be different from the usual formula for variance of the least squares slope estimator when an intercept is present.]

(9.5)) [Fundamental assumptions] Consider the two-variable model *without* a slope, that is $y = \beta + \varepsilon$. Notice that there is only one β parameter for this model—the intercept. There is no intercept because the true population regression function is here assumed to be horizontal. It can be shown that the least-squares estimator of β is given by $\hat{\beta} = \frac{1}{n}\sum y_i$. where the sum runs from i=1 to n. Assume that $E(\varepsilon_i) = 0$.

a. Prove that $\hat{\beta} = \beta + \frac{1}{n}\sum \varepsilon_i$. Justify each step of your proof.

b. Prove that the least-squares estimator of β is unbiased, that is $E(\hat{\beta}) = \beta$. Justify each step of your proof.

c. Further assume that $Var(\varepsilon_i) = \sigma^2$ and that $Cov(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$. Derive the formula for the variance of the least-squares estimator $Var(\hat{\beta}) = E\left(\left(\hat{\beta} - \beta\right)^2\right)$.

   Justify each step of your proof. [Hint: Your formula will be different from the usual formula for variance of the least squares estimator when a slope is present.]

(9.6) [Fundamental assumptions] Suppose we wish to estimate the effect of education on wages *ceteris paribus*, using data on workers. We use the regression equation $w_i = \beta_1 + \beta_2 s_i + \varepsilon_i$, where s_i denotes the number of years of school completed, and w_i denotes hourly wages. We presume $\beta_2 > 0$, that is, education enhances wages, so the true line slopes upward.

a. What other characteristic of workers (other than the number of years of school) might affect their wages? Would it have a *positive* or a *negative* effect on wages?

b. Would you expect that other variable to be *positively* correlated, *negatively* correlated or *uncorrelated* with x, the number of years of school completed? Why?

c. Since our equation does not include that other variable explicitly, it must be part of the error term ε. Given your answer to part (b), which of the Gauss-Markov assumptions must be violated in this situation? Explain why, using a sketch if possible.

d. Will the least-squares estimator $\hat{\beta}_2$ be biased *upward, downward, or not at all?* That is, will the least-squares estimator $\hat{\beta}_2$ tend to be larger or smaller than the true value β_2? Explain why, using a sketch if possible.

(9.7) [Fundamental assumptions] Suppose we wish to estimate the effect of policing on the crime rate *ceteris paribus*, using data on cities. We use the regression equation $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, where x_i denotes the number of police officers per thousand population, and y_i denotes serious crimes per thousand population. We presume $\beta_2 < 0$, that is, policing deters crime, so the true line slopes downward.

a. What other characteristic of cities (other than the number of police officers) might affect their crime rates? Would it have a *positive* or a *negative* effect on crime rates?
b. Would you expect that other variable to be *positively* correlated, *negatively* correlated or *uncorrelated* with x, the number of police officers? Why?
c. Since our equation does not include that other variable explicitly, it must be part of the error term ε. Given your answer to part (b), which of the Gauss-Markov assumptions must be violated in this situation? Explain why, using a sketch if possible.
d. Will the least-squares estimator $\hat{\beta}_2$ be biased *upward, downward, or not at all?* That is, will the least-squares estimator $\hat{\beta}_2$ tend to be more negative or less negative than the true value β_2? Explain why, using a sketch if possible.

(9.8) [Additional assumptions] Suppose we estimate the model $y = \beta_1 + \beta_2 x + \varepsilon$ by least-squares using a sample of n=100 observations. Our initial calculations yield the following.

$$\sum (x_i - \bar{x})^2 = 256. \quad \text{and} \quad \sum \hat{\varepsilon}_i^2 = 2450.$$

Assume all four basic least-squares assumptions hold: the error term has mean zero, the error term is uncorrelated with x, the error term is homoskedastic, and there is no autocorrelation.

a. Compute the unbiased estimate of the variance of the error term.
b. Compute the standard error of the least-squares slope estimator, $\hat{\beta}_2$.

(9.9) [Confidence intervals, t-tests] Suppose the relationship between household income and electricity usage is modeled as a linear relationship $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, where y denotes annual electricity use (in kilowatt hours) and x denotes annual household income in dollars. This equation is estimated by least-squares on a sample of several thousand households, with the following results. The numbers on top are the least-squares estimates of the intercept and slope, and the numbers in parentheses are standard errors.

Electricity usage in kilowatt-hours	=	2149.0 (614.0)	+	0.58 (0.25)	Household income in dollars

a. Compute a 95% asymptotic confidence interval for the intercept.

Test the null hypothesis that household income has no effect on electricity usage, against the alternative hypothesis that income has a positive effect (a one-tailed test) at 5% significance.

b. Give the value of the test statistic.
c. Give the critical point from the appropriate table at the back of your textbook (or compute the p-value using a spreadsheet program).
e. Give your conclusion: whether you reject the null hypothesis at 5% significance.

(9.10) [Confidence intervals, t-tests] Suppose the relationship between college GPA and high school GPA is modeled as a linear relationship $CGPA_i = \beta_1 + \beta_2 HSGPA_i + \varepsilon_i$. This equation is estimated by least-squares on a sample of several thousand college sophomores, with the following results. The numbers on top are the least-squares estimates of the intercept and slope, and the numbers in parentheses are standard errors.

CGPA	=	1.02 (0.24)	+	0.68 (0.16)	HSGPA

a. Compute a 95% asymptotic confidence interval for the intercept.

Test the null hypothesis that high school GPA has no effect on college GPA, against the alternative hypothesis that high school GPA has some effect (a two-tailed test) at 5% significance.

b. Give the value of the test statistic.
c. Give the critical point from the appropriate table at the back of your textbook (or compute the p-value using a spreadsheet program).
e. Give your conclusion: whether you reject the null hypothesis at 5% significance.

[end of problem set]